From FRE to Curator

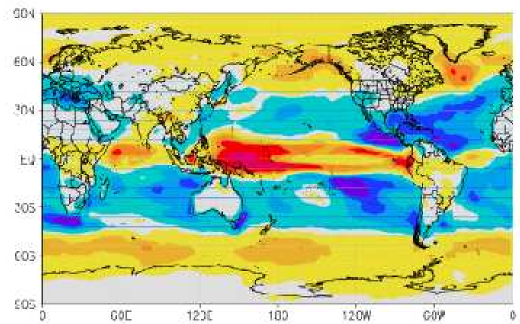
V. Balaji
Princeton University

22 September 2006

The IPCC AR4 archive at PCMDI

The IPCC data archive at PCMDI is a truly remarkable resource for the comparative study of models. Since it came online in early 2005, it has been a resource for ∼300 scientific papers aimed at providing consensus and uncertainty estimates of climate change, from ∼20 state-of-the-art climate models worldwide.

Model	Modeling Center
BCCR BCM2	Bjerknes Centre for Climate Research
CCCMA CGCM3	Canadian Centre for Climate Modeling & Analysis
CNRM CM3	Centre National de Recherches Meteorologiques
CSIRO MK3	CSIRO Atmospheric Research
GFDL CM2_0	Geophysical Fluid Dynamics Laboratory
GFDL CM2_1	Geophysical Fluid Dynamics Laboratory
GISS AOM	Goddard Institute for Space Studies
GISS EH	Goddard Institute for Space Studies
GISS ER	Goddard Institute for Space Studies
IAP FGOALS1	Institute for Atmospheric Physics
INM CM3	Institute for Numerical Mathematics
IPSL CM4	Institut Pierre Simon Laplace
MIROC HIRES	Center for Climate System Research
MIROC MEDRES	Center for Climate System Research
MIUB ECHO	Meteorological Institute University of Bonn
MPI ECHAM5	Max Planck Institute for Meteorology
MRI CGCM2	Meteorological Research Institute
NCAR CCSM3	National Center for Atmospheric Research
NCAR PCM1	National Center for Atmospheric Research
UKMO HADCM3	Hadley Centre for Climate Prediction



This figure, from Held and Soden (2005), is a composite analysis across the entire IPCC archive.

Computational load at GFDL:

- 5500 model years run.
- Occupied half of available compute cycles at GFDL for half a year (roughly equivalent to 1000 Altix processors).
- 200 Tb internal archive; 40 Tb archived at GFDL data portal; 4 Tb archived at PCMDI data portal.

I would argue that the IPCC experiment is ***already*** petascale!

The FMS user interface: FRE

Comprehensive website for all information and documentation:
`http://www.gfdl.noaa.gov/~fms`

- Source code maintenance under CVS; browse over the net using webCVS.
- Model configuration, launching and regression testing encapsulated in XML;
- Relational database for archived model results;
- Standard and custom diagnostic suites;

The FMS Runtime Environment (FRE) describes all the steps for configuring and running a model jobstream; archiving, postprocessing and analysis of model results.

`fremake, frerun, frepp, frecheck, ...`

The Regression Test Suite (RTS) is a set of tests that are run continuously on a set of FMS models to maintain and verify code integrity.

Elements of FRE

fremake Checkout an appropriate subset of the FMS source code for an experiment and create an executable;

frerun run an experiment in multiple ***segments***; resubmit if necessary;

frestatus check the status of an experiment that is underway;

frelist list available experiments;

frepriority switch a job sequence between queues;

frecheck run RTS checks for bitwise accuracy;

frepp FRE post-processing: create time series, time averages, and plots;

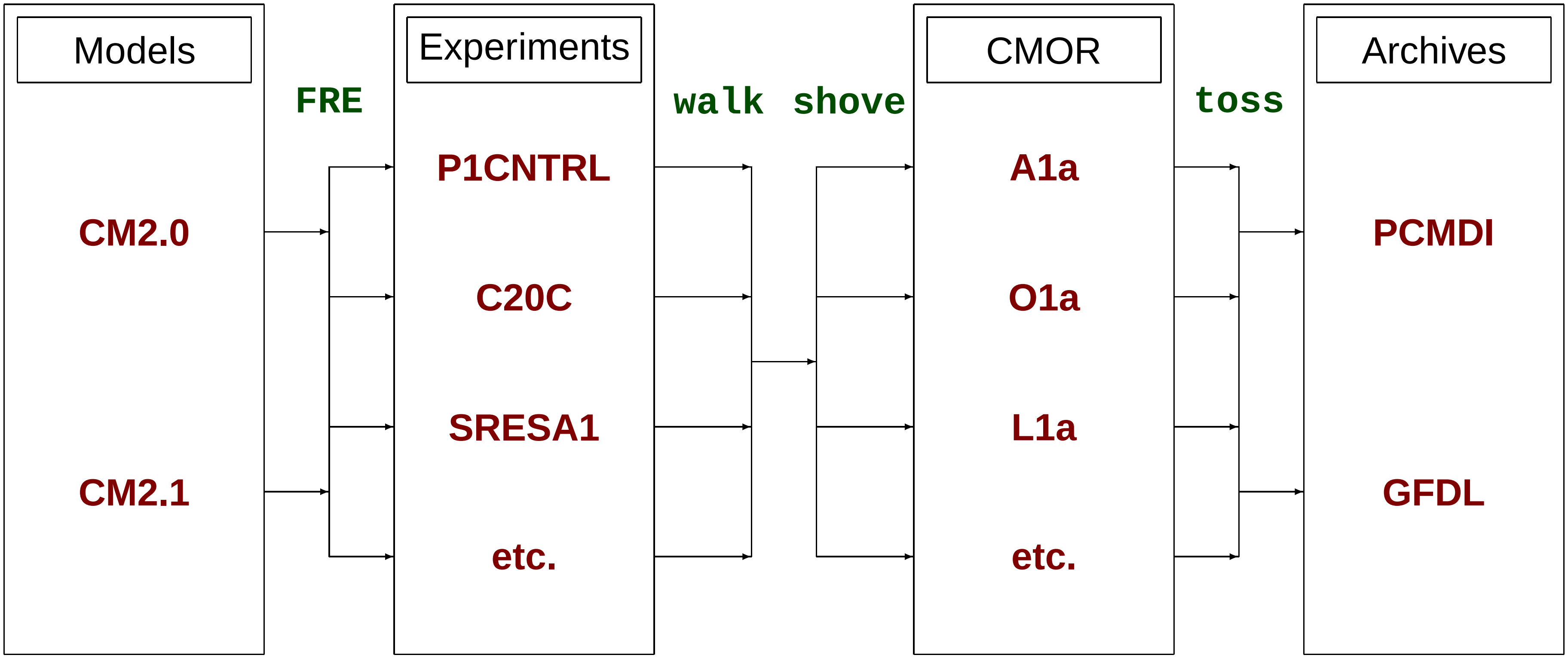
frescrub remove intermediate and redundant files;

freppcheck RTS checks on history and post-processing files.

freversion tool to upgrade the XML, should the syntax change.

URL: **`http://www.gfdl.noaa.gov/fms/fre`**

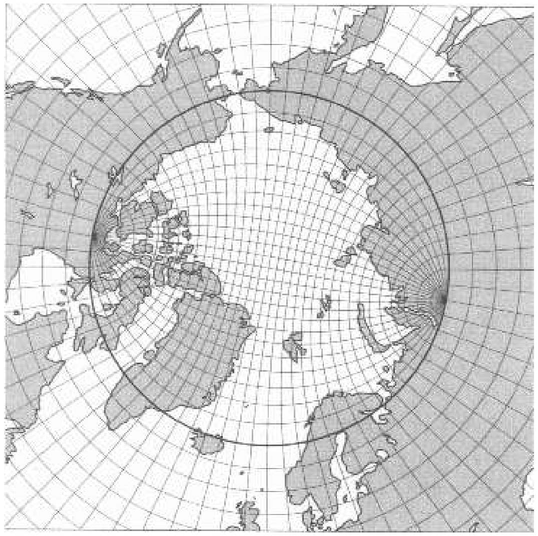
The IPCC data pipeline at GFDL

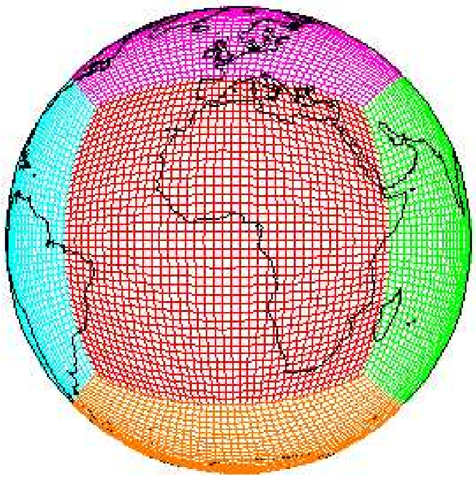


The process was time- and data-intensive, with multiple access episodes for the same datasets. Clearly it would be ideal if FRE already produced compliant data.

Current problems with CMOR-compliant data

- A principal difficulty is CMOR's restricted view of model grids: only simple latitude-longitude grids are permitted. This is because the current crop of visualization and analysis tools cannot easily translate data among different grids. Shown at right are the ***tripolar grid*** (Murray 1996, Griffies et al 2004) used by MOM4 for GFDL's current IPCC model CM2. Below is the ***cubed sphere*** (Rancic and Purser 1990) planned for the Finite-Volume atmosphere dynamical core for the next-generation GFDL models AM3 and CM3. If there were a ***grid metadata standard***, regridding operations could potentially be applied by the end-user using standard-compliant tools.
- The model descriptions demanded by CMOR do not contain enough information about the models, and are added after the fact. If there were a ***model metadata standard*** such as NMM in force, comprehensive model descriptions could be automatically produced. The end-user could better diagnose specific differences between different models in an archive.

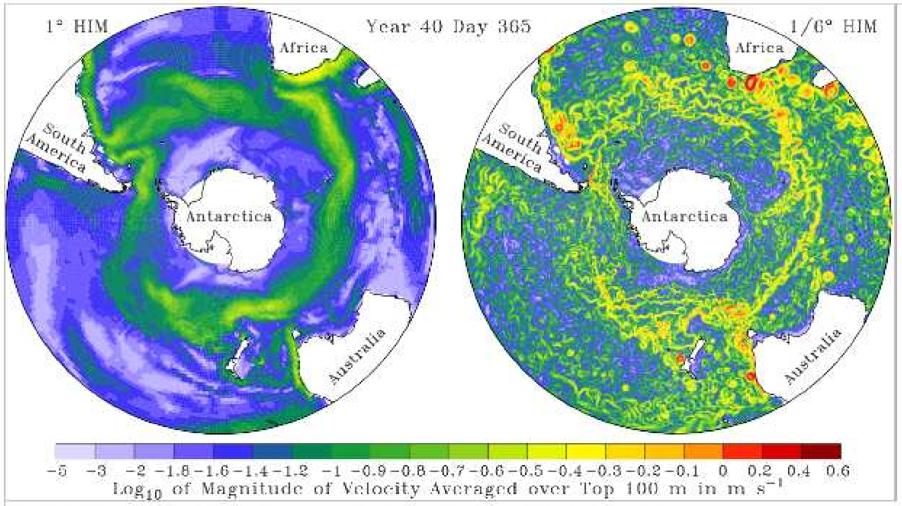




Can an experiment like IPCC be run at higher resolution?

Possible key challenges for the next IPCC:

- Robust estimates of regional climate change.
- Interactive carbon dynamics: inclusion of land-use change, ocean carbon uptake, marine and terrestrial biospheric response to climate change.
- Increased resolution in the atmosphere (even before we get to cloud-resolving scales) will lead to better characterization of storm track changes and hurricane intensity projections in a changed climate. Target: 1°or 0.5°model for IPCC AR5.
- Increased resolution in the ocean is even more critical: key mechanisms of ocean mass and energy transport are currently unresolved. Targets: 0.25°("eddy-permitting") models next time around, 0.0825°("eddy-resolving") still out of reach.



Shaded is the speed of the surface currents in the ocean model at two resolutions.

Petascale methodologies

As much emphasis must be placed on methodologies to facilitate scientific analysis of multi-model ensembles on distributed data archives, as on the computational technology itself.

Some current efforts:

ESC Earth System Curator, funded by NSF. Partners GFDL, NCAR, PCMDI, Georgia Tech. Will be used to promote the existence of a model and grid metadata standard, and build a prototype relational database housing these metadata. Will build tools for model configuration and compatibility checking based on automatic harvesting of metadata from code.

MAPS Modeling, Analysis and Prediction System? funded by NASA, partners NASA/GSFC, GFDL, MIT. Proposes to build a configuration layer for a subset of coupled models based on PRISM config files, and conformant with grid and metadata standards. Will attempt to promote a "standard coupling architecture" and develop a standard for exchange grids for ESMF.

GO-ESSP and CF should be the medium of exchange for standard-building. CF is seeking funding and WGCM backing to become a mandated activity. GO-ESSP is the ideal medium for the actual technical work of standard-building.

IPCC! PCMDI and other data centres should be core participants.

With a complete metadata hierarchy defined, one can envisage the convergence of modeling and data frameworks into a single environment: a model ***curator***.

Scenario 1: dynamically generated data catalogues

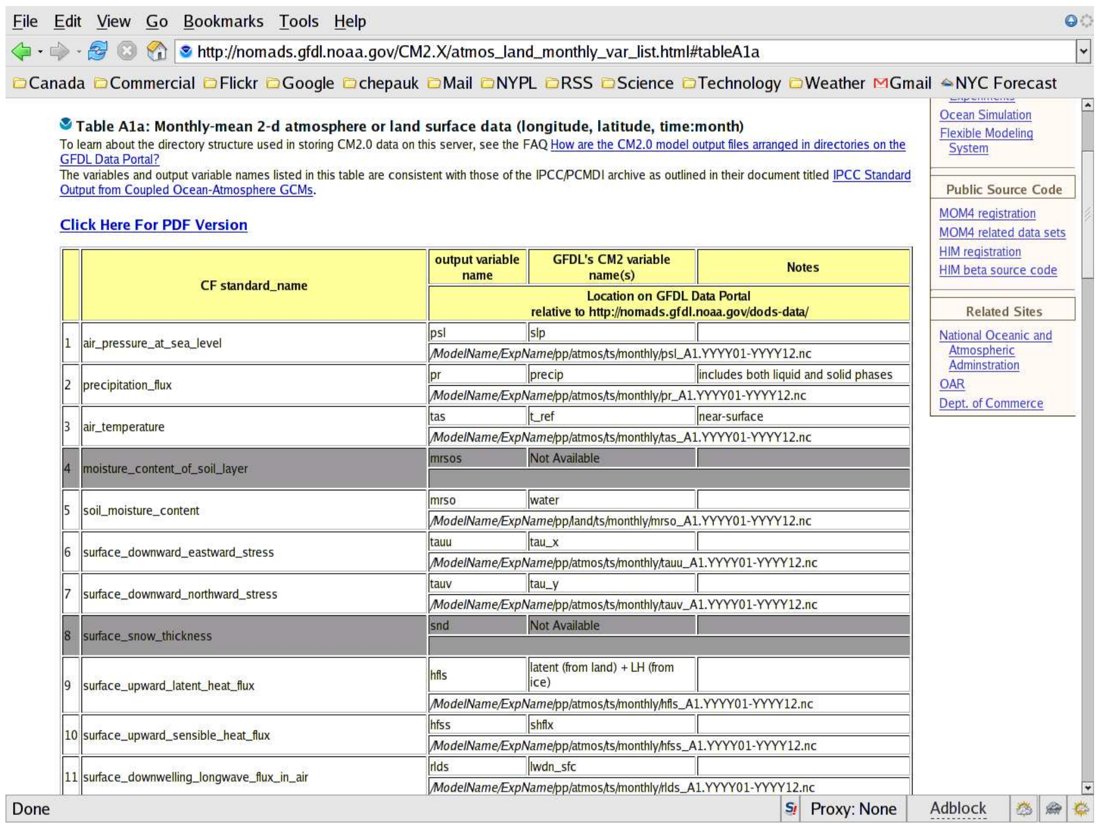


Table A1a: Monthly-mean 2-d atmosphere or land surface data (longitude, latitude, time:month)

To learn about the directory structure used in storing CM2.0 data on this server, see the FAQ How are the CM2.0 model output files arranged in directories on the GFDL Data Portal?

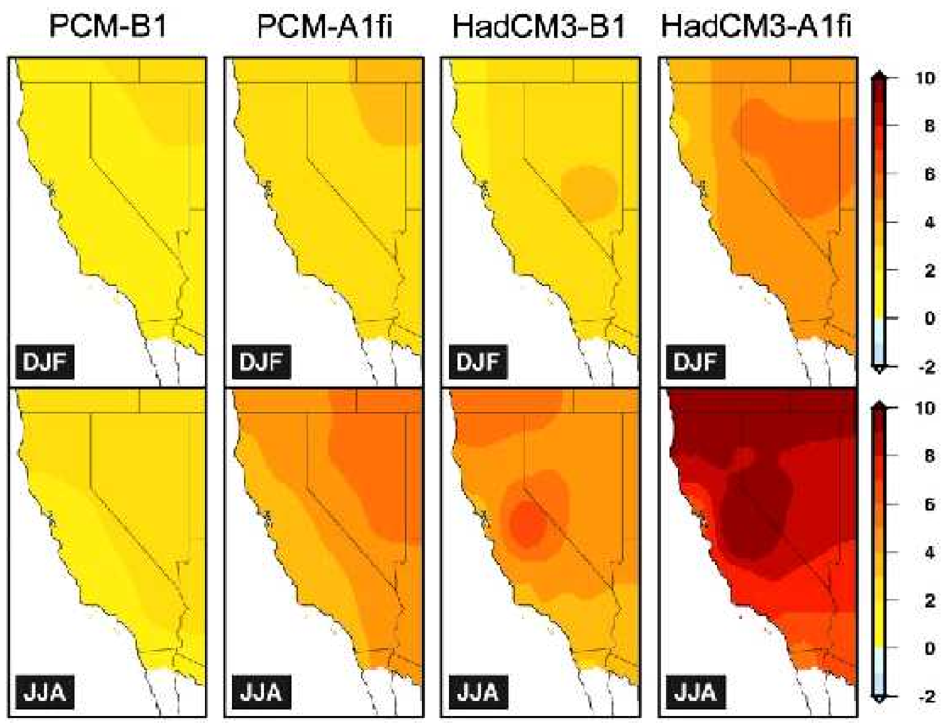
The variables and output variable names listed in this table are consistent with those of the IPCC/PCMDI archive as outlined in their document titled IPCC Standard Output from Coupled Ocean-Atmosphere GCMs.

Click Here For PDF Version

	CF standard_name	output variable name	GFDL's CM2 variable name(s)	Notes
		Location on GFDL Data Portal relative to http://nomads.gfdl.noaa.gov/dods-data/		
1	air_pressure_at_sea_level	psl	slp	
		/ModelName/ExpName/pp/atmos/ts/monthly/psl_A1.YYYY01-YYYY12.nc		
2	precipitation_flux	pr	precip	includes both liquid and solid phases
		/ModelName/ExpName/pp/atmos/ts/monthly/pr_A1.YYYY01-YYYY12.nc		
3	air_temperature	tas	t_ref	near-surface
		/ModelName/ExpName/pp/atmos/ts/monthly/tas_A1.YYYY01-YYYY12.nc		
4	moisture_content_of_soil_layer	mrsos	Not Available	
5	soil_moisture_content	mrso	water	
		/ModelName/ExpName/pp/land/ts/monthly/mrso_A1.YYYY01-YYYY12.nc		
6	surface_downward_eastward_stress	tauu	tau_x	
		/ModelName/ExpName/pp/atmos/ts/monthly/tauu_A1.YYYY01-YYYY12.nc		
7	surface_downward_northward_stress	tauv	tau_y	
		/ModelName/ExpName/pp/atmos/ts/monthly/tauv_A1.YYYY01-YYYY12.nc		
8	surface_snow_thickness	snd	Not Available	
9	surface_upward_latent_heat_flux	hfls	latent (from land) + LH (from ice)	
		/ModelName/ExpName/pp/atmos/ts/monthly/hfls_A1.YYYY01-YYYY12.nc		
10	surface_upward_sensible_heat_flux	hfss	shflx	
		/ModelName/ExpName/pp/atmos/ts/monthly/hfss_A1.YYYY01-YYYY12.nc		
11	surface_downwelling_longwave_flux_in_air	rlds	lwdn_sfc	
		/ModelName/ExpName/pp/atmos/ts/monthly/rlds_A1.YYYY01-YYYY12.nc		

Already in use at PCMDI, DDC, GFDL Curator, elsewhere: metadata requires extension.

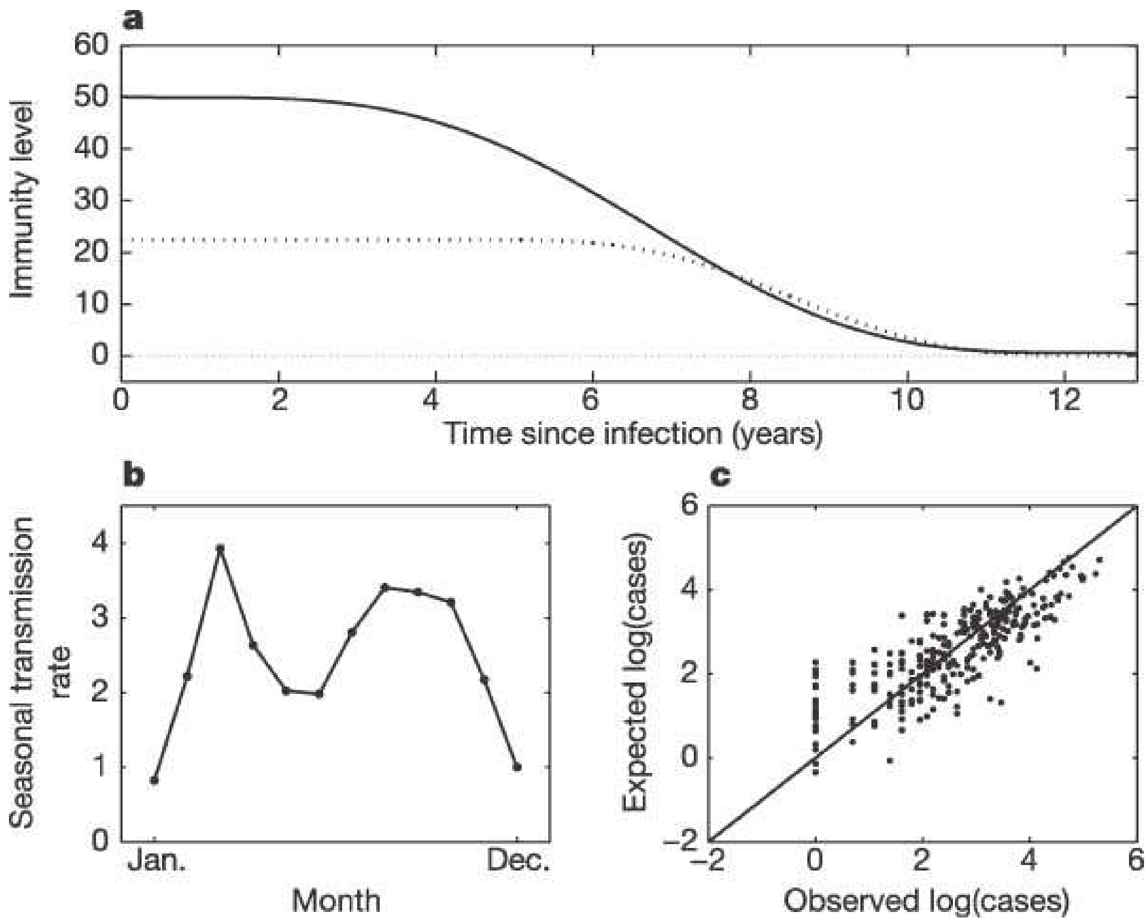
Scenario 2: statistical downscaling of climate change projections



Hayhoe et al, ***PNAS***, 2004: ***Emissions pathways, climate change, and impacts on California.***

Uses daily data for "heat degree days" and other derived quantities. Requires data beyond that provided by IPCC AR4 SOPs (1960-2000).

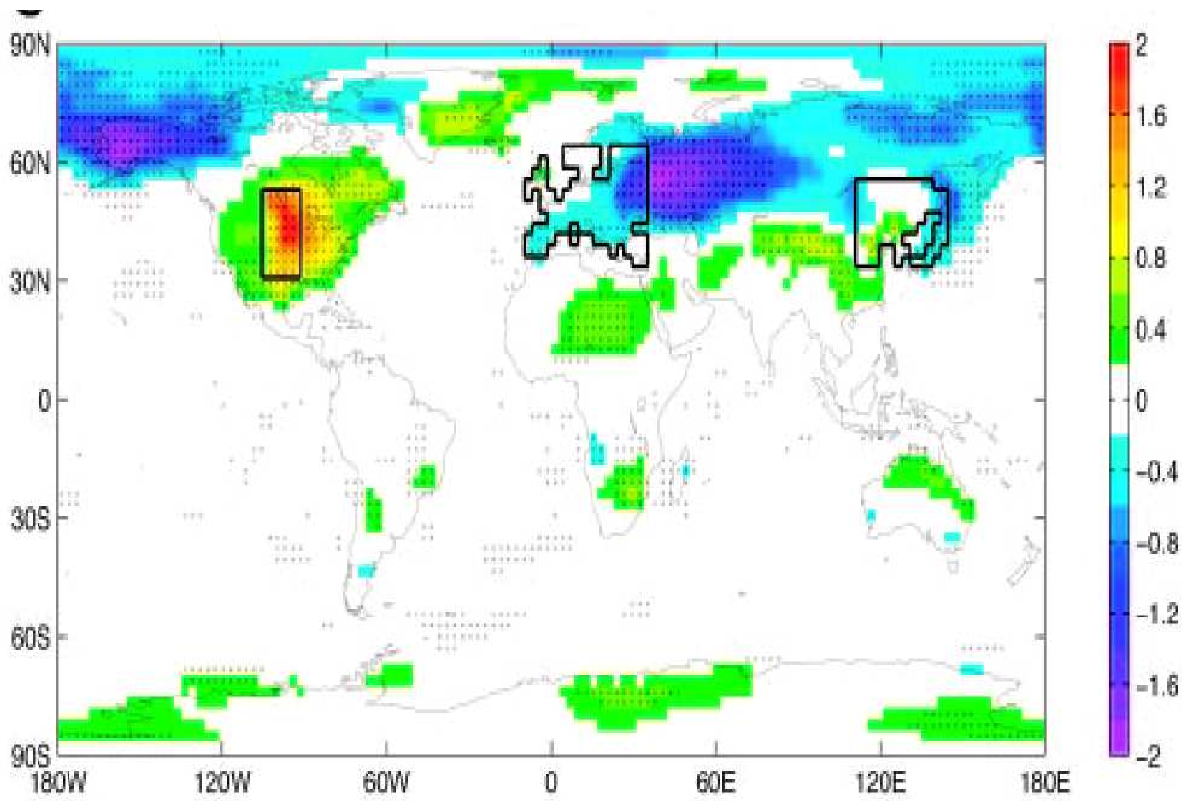
Scenario 3: disease vectors in a changing climate



Koelle et al, ***Nature***, 2005: ***Refractory periods and climate forcing in cholera dynamics.***

Requires monthly forcing data, no feedback.

Scenario 4: alternate energy sources



Keith et al, ***PNAS***, 2005: ***The influence of large-scale wind power on global climate.***

Feedback on atmospheric timescales: but does not require model to be retuned.

Taking stock halfway through the noughties

- Earth system models are evolving into powerful tools for advancing our understanding, and well on their way to being operational tools in support of policy and industrial strategy.

- The principal research path for consensus and uncertainty estimates of climate change is the comparative study of models.

- The building of appropriate standards has been identified as a key element in uniting modeling and data communities.

- This requires convergence and cross-fertilization between model and data frameworks: by developing a clear understanding of the architecture of Earth system models, PRISM and ESMF also point the way to a metadata hierarchy to be used in building curators.

- Leadership in standards will come from custodians of international multi-model data archives well connected to data consumers, and will be embedded in the modeling frameworks.

- Research is needed into hierarchical data storage, use of pattern recognition and feature detection for data reduction, remote data analysis and visualization.